UNIDAD 1 “ESTUDIO DESCRIPTIVO DE DATOS”

1.1 Definiciones básicas

Datos: Son hechos, números o información que se recogen, analizan y resumen para ser interpretados y presentados.

Tipos de datos:

- *Cualitativos*: Representan características, cualidades o categorías.
  - Nominal: categorías sin orden (género, color de ojos, marca de auto).
  - Ordinal: categorías con orden (nivel socioeconómico: bajo, medio, alto).
- *Cuantitativos*: requieren valores numéricos, se pueden medir o contar. Se obtienen usando las escalas de medición de intervalo o de razón.

Elemento: Cada objeto o individuo del que se obtienen datos.

Variable: Característica que se mide en los elementos.

Observación: Conjunto de mediciones obtenidas para cada elemento

Tipos de variables:

- *Cualitativas*: describen cualidades, categorías o características
- *Cuantitativas*: son numéricos, se pueden medir o contar.
  - Discretas: toman valores enteros (número de hijos, número de autos)
  - Continuas: toman cualquier valor dentro de un intervalo específico (peso, tiempo, temperatura).
- **Datos transversales**: obtenido en el mismo o aproximadamente el mismo momento
- **Datos de series de tiempo**: recogidos a lo largo del tiempo. Obtenidos a lo largo de varios periodos.

Estadística: Reunir, organizar, resumir, presentar, analizar e interpretar datos para tomar decisiones informadas.

- *Estadística descriptiva*: recopilación y descripción de un grupo de datos. (tablas, histogramas, medias, desviaciones).

- *Estadística inferencial*: métodos para tomar decisiones / generalizaciones sobre una población a partir de una muestra, usando probabilidad.

Población (o universo): Conjunto completo de todos los elementos que nos interesan en un estudio.

Muestra: Subconjunto de la población, seleccionado para el estudio. Porción o parte de la población de interés.

1.2 Histograma de frecuencias relativas

Una **tabla de frecuencias** organiza los datos en clases, indicando cuántas veces aparece cada valor.

- *Datos agrupados*: datos organizados en una distribución de frecuencia.
- *Datos no agrupados*: se enumeran todos los valores observados de la variable aleatoria.

Un **histograma** es un gráfico de barras que representa una distribución de frecuencias:

- Eje X: clases. (El número de clases suele estar entre 5 y 15)
- Eje Y: frecuencias (o frecuencias relativas).

Elementos de una distribución:

- **Frecuencia absoluta (f):** número de veces que aparece un valor o intervalo.
- **Frecuencia relativa (fr):** proporción de veces que aparece, respecto al total.
- **Frecuencia acumulada:** suma progresiva de las frecuencias.
- **Punto medio de clase:** promedio de los límites inferior y superior de cada clase.

Organización de los datos:

- **Serie simple**: datos ordenados de menor a mayor sin agrupar.
- **Serie de clases y frecuencias**: datos agrupados en clases con frecuencias.
- **Serie de clases por intervalos**: clases definidas por límites inferiores y superiores.

El histograma nos ayuda a ver la forma de la distribución: simétrica, sesgada, con más de un pico, etc.

Polígono de frecuencias: gráfico lineal que une los puntos correspondientes a las frecuencias de cada clase.

1.3 Medidas de tendencia central

Son valores que representan el "centro" o el valor típico de los datos.

Media aritmética (promedio): Se obtiene sumando todos los valores y dividiendo entre el número de datos.

Ejemplo: notas = 8, 7, 9. Media = (8+7+9)/3 = 8.

Mediana: Valor que ocupa el lugar central cuando los datos están ordenados de mayor a menor o viceversa.

- Si hay número impar de datos → el dato central.
- Si hay número par → promedio de los dos datos centrales.

Moda: Es el valor que más se repite.

1.4 Medidas de variabilidad (dispersión)

Indican qué tan dispersos o concentrados están los datos alrededor de la media.

Media aritmética

- Datos no agrupados:

POBLACIÓN $$\mu = \frac{\Sigma X}{N}$$

MUESTRA $$\bar{X} = \frac{\Sigma X}{n}$$

- Datos agrupados: POBLACIÓN $\mu = \frac{\Sigma(fX)}{N}$ MUESTRA $\bar{X} = \frac{\Sigma(fX)}{n}$

Mediana

- Datos no agrupados: $\text{Med} = X_{[(n/2)+(1/2)]}$

- Datos agrupados: $\text{Med} = L_I + \left(\frac{\frac{N}{2} - fa_A}{f_c} \right) i$

Moda

- Datos no agrupados: La moda para estos datos es el valor que tiene la mayor frecuencia.

- Datos agrupados: $\text{Moda} = L_I + \left(\frac{d_1}{d_1 + d_2} \right) i$

Desviación media: Es el promedio aritmético de los valores absolutos de las desviaciones con respecto a la media aritmética o promedio.

- Datos no agrupados:

$$\text{DM poblacional} = \frac{\Sigma |X - \mu|}{N}$$

$$\text{DM muestral} = \frac{\Sigma |X - \bar{X}|}{n}$$

- Datos agrupados:

$$\text{DM poblacional} = \frac{\Sigma(f|X - \mu|)}{N}$$

$$\text{DM muestral} = \frac{\Sigma(f|X - \bar{X}|)}{n}$$

Varianza: Es el promedio de las diferencias al cuadrado entre cada dato y la media.

- Datos no agrupados:

POBLACIÓN $$\upsilon(X) = \sigma^2 = \frac{\Sigma(X) - \mu)^2}{N}$$

MUESTRA $$s^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$$

- Datos agrupados:

POBLACIÓN $$\sigma^2 = \frac{\Sigma[f(X-\mu)^2]}{N}$$

MUESTRA $$s^2 = \frac{\Sigma[f(X-\bar{X})^2]}{n-1}$$

Desviación estándar: Es la raíz cuadrada positiva de la varianza. Mientras más pequeña, más homogéneos los datos.

- Datos no agrupados:

POBLACIÓN $$\sigma = \sqrt{\frac{\Sigma X^2 - N\mu^2}{N}}$$

MUESTRA $$s = \sqrt{\frac{\Sigma X^2 - n\bar{X}^2}{n-1}}$$

- Datos agrupados:

POBLACIÓN $$\sigma = \sqrt{\frac{\Sigma[f(X-\mu)^2]}{N}}$$

MUESTRA $$s = \sqrt{\frac{\Sigma[f(X-\bar{X})^2]}{n-1}}$$

1.5 Medidas de posición

Muestran dónde se encuentra un valor dentro de la distribución de los datos.

Cuartiles (Q1, Q2, Q3): dividen los datos en 4 partes iguales.

- Q1 = 25% de los datos están debajo.
- Q2 = 50% (equivale a la mediana).
- Q3 = 75% de los datos están debajo.

Deciles: dividen los datos en 10 partes iguales.

Percentiles: dividen los datos en 100 partes iguales.

- Datos no agrupados:

$$Q_1 \text{ (primer cuartil)} = X_{[(n/4)+(1/2)]}$$

$$D_3 \text{ (tercer decil)} = X_{[(3n/10)+(1/2)]}$$

$$P_{70} \text{ (percentil 70)} = X_{[(70n/100)+(1/2)]}$$

- Datos agrupados:

$$Q_1 \text{ (primer cuartil} = L_I + \left(\frac{\frac{n}{4} - fa_A}{f_c} \right) i$$

$$D_3 \text{ (tercer decil)} = L_I + \left(\frac{\frac{3n}{10} - fa_A}{f_c} \right) i$$

$$P_{70} \text{ (percentil 70)} = L_I + \left(\frac{\frac{70n}{100} - fa_A}{f_c} \right) iI$$